# Results of the Collaborative Energy and Water Cycle Information Services (CEWIS) Workshop on Heterogeneous Dataset Analysis Preparation

Steven Kempler[1], William Teng[2], James Acker[2], Deborah Belvedere[3], Zhong Liu[4], Gregory Leptoukh[1]

[1]NASA Goddard Space Flight Center, [2]NASA Goddard Space Flight Center/SESDA2, [3]UMBC/GEST, [4]GMU

*Steven.J.Kempler@nasa.gov*

## Abstract

In support of the NASA Energy and Water Cycle Study (NEWS), the Collaborative Energy and Water Cycle Information Services (CEWIS), sponsored by NEWS Program Manager Jared Entin, was initiated to develop an evolving set of community-based data and information services that would facilitate users to locate, access, and bring together multiple distributed heterogeneous energy and water cycle datasets. The CEWIS workshop, June 15-16, 2010, at NASA/GSFC, was the initial step of the process, starting with identifying and scoping the issues, as defined by the community.

## Motivation

•Cross-dataset analysis is growing, requiring much effort for data preparation, by EACH researcher
•Difficulty in locating and using heterogeneous datasets together for global and regional energy and hydrology research
•Datasets need to contain uniform characteristics
•New technologies are available to mitigate data inter-comparison issues
•Facilitate cross dataset data validation

## Challenges in Performing Multi-Source Data Inter-comparisons

- Data are distributed, heterogeneous (data formats, structure), high volume
- Requires data specific software
- Requires homogeneous data analysis, management, visualization services
- **Data sets are often unprepared for open availability and interoperability**

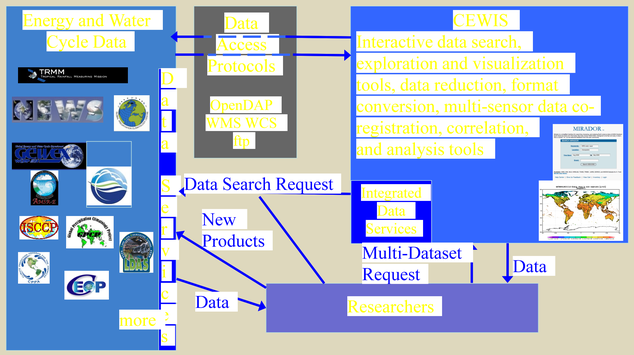
## Workshop Results

1.Participant responses to survey questions pertaining to interoperability of heterogeneous datasets.
2.Participant presentations that provide real multi-dataset research preparation experiences.
3.Participant developed multi-dataset preparation scenarios.
4.Discussion at the end of the meeting providing insights on where to go from here.

## It is hoped…

… that this presentation will encourage further discussions and collaborations on behalf of promising information technologies that would facilitate the preparation of heterogeneous datasets for science and applications research.

## The Collaborative Effort Model




## Workshop Results (http://news.cisc.gmu.edu/cewisworkshop.htm)

### 1. Survey Responses

**1. What steps do you take to gather and prepare data so that you can perform multi-data set inter-comparisons?**

***Responses:***

Retrieving Data
- Find relevant data with the desired data characteristics
- Get samples of data; sort out data 'quirks'; determine if data are readable/correct
- Check units, timestamp, quality control flags
- Understand data characteristics (format, time period and resolution)

Assembling Data
- Collocate from different instruments
- Bring data sets to common grid (interpolate, collocate)

Analyzing Data
- Acquire data read code
- Perform data subsetting
- Perform data inter-comparisons
- Homogenize different data sets for objective comparisons

**2. What data-related roadblocks do you encounter when bringing heterogeneous data sets together?**

***Responses:***

Data Access
- Finding and gaining access to the data
- Data sets tend to be organized on a project-specific basis
- Lack of a nice "search engine" to quickly locate the data.

Data Characteristics
- Learning how to read data correctly
- Data volume → download interruption
- Spatially and temporally subsetting data

Combining Datasets
- Finding collocated data sets
- Converting data of different formats
- Properly converting data to common grid
- Users are expected to have knowledge of multiple sensors
- Dealing with different spatial/temporal resolutions, spanning periods, aspects of heterogeneous data sets, instrument fields of view

Verifying Combined Data sets
- Quantifying errors introduced during interpolation
- Identifying natural signal from systematic errors
- Time/space gridding mismatches across data sets

Data Documentation
- Undocumented features in data
- Inadequate quality control, error estimation, detailed documentation of data

### 2. Multi-Dataset Research Preparation Experiences

**Four presentations were given relating experiences in acquiring and utilizing heterogeneous datasets scientific research:**

•**Multi-Dataset Collection Research Scenario – GPCP and TMPA -** George J. Huffman, David T. Bolvin, and Eric J. Nelkin
•**Multi-Dataset Collection Research Scenario – Air France Flight 447 Case Study -** Zhong Liu, Dana Ostrenga, and Greg Leptoukh
•**Merged Atmospheric Water Data Set from A-Train and A Multi-Sensor Water Vapor Climate Data record Using Cloud Classification** - Eric Fetzer
•**The Merged Surface and Satellite Observed Cloud, Radiation, and Precipitation Data Sets** - Baike Xi, X. Dong, Z. Feng, A. Kennedy, T. Longan, B. Zib, D. Wu, K. Giannecchini and Y. Qiu

**Their collective experiences in regards to collecting and utilizing heterogeneous data include:**

- Hear about the data and find data
- Get data samples
- Modify/adapt/build reading code
- (Re)grid
- Retrieve high volume data
- Learn formats and develop readers
- Extract parameters
- Perform spatial subsetting
- Perform filtering
- Identify quality flags and other dataset caveats
- Develop analysis and visualization tools
- Accept/discard/get more data
- Assembling data sets
- Collocating observations from different instruments
- Reconciling and cross-linking L1 & L2 observations
- "90% of the time is spent getting data into the right format"

### 3. Multi-Dataset Preparation Scenarios

1. Processing a global land flux data set at NCCS
2. Bring NEWS data sets having disparate characteristics together
3. Routine data assimilation (NLDAS, GLDAS Drought Monitoring)
4. Comprehensive, cross-discipline datasets for events or seasons
5. Make point data representational of an area
6. Assembly of data set from multiple instruments/satellite and variables
7. Search capabilities using time and location and variable
8. Include non-satellite data when relevant
9. Creating match up data sets
10. Provide analysis tools
11. Providing data to non expert users (L3 /gridded data)

### 4. Participant Discussion

**Participant notes and discussion lead to the following workshop conclusions:**

•Detailed descriptions of each dataset is very important for combining heterogeneous datasets
•There is a need by the broader community for tools to facilitate combining heterogeneous datasets
•Deficiencies in dataset preparation make data potentially difficult to use
•Simple visualization and analysis tools, are needed, with the capability to do intercomparisons, spatial plotting, time series plots of the mean and standard deviation of a specified region, scatter plot

**Suggested for short term efforts:**

•Provide information on historical usage of data sets (i.e., provenance)
•Distinguish process vs. climatological studies; non-gridded/non-averaged vs. gridded/averaged data
•OpenSearch results list is too long
•Define and implement data co-registration (e.g., pixel matching, common grid)
•Data centers should be registering data sets
•Emphasis was put on the need for usage guidelines, documentation, context-sensitive information
•Need for standards on data description, documentation, terminology
•Strong need to have data access latency for remote data as if on local disk
•Provide metrics on data usage

**Longer term questions/concepts:**

•How can data documentation be maintained and improved?
•Organize datasets in a way that addresses the NEWS community (engage the community)
•Hold a 2nd workshop, to address available community wide multi-dataset research tools and services, specifically focusing on addressing the scenarios, issues, and roadblocks discussed in this workshop
•Report Workshop findings at a NEWS PI meeting
•Address data harmonization: QA, formats, resolution etc.
•Bring together community members who are already having success overcoming the problems we have discussed
•Have PI's provide prioritized lists of data sets needed for multi dataset studies.
•Engage broader energy and water cycle community
•Further expand data access services
•Formulate a Community Advisory Committee to seek community consensus on desirable tools and CEWIS type functionality direction
•Establish a series of mutually beneficial data and service exchanges with pertinent projects (e.g., GEWEX, NEWS, CEOP)